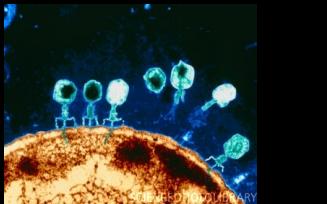


Introducción a la Inferencia Filogenética – las especies y genomas microbianos

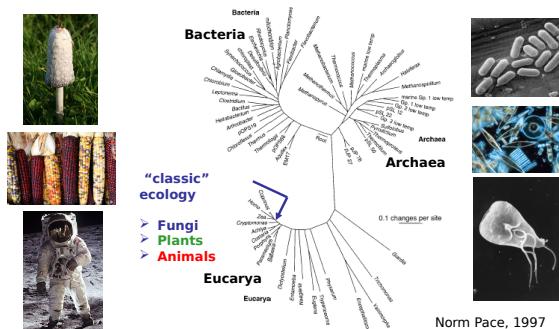
Conceptos de filo-informática para investigación en genómica, ecología y evolución microbiana

Pablo Vinuesa, Centro de Ciencias Genómicas – UNAM
vinuesa [at] ccg . unam . mx
<http://www.ccg.unam.mx/~vinuesa/>



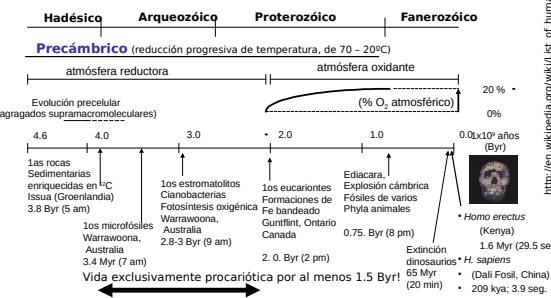
Biodiversity = microbial diversity

- the SSU rDNA view of diversity on Earth



Evolución orgánica - La dimensión temporal

Historia de la tierra y de la vida



Parte I: Introducción a la Inferencia Filogenética

Conceptos básicos:

* filogenia, evolución molecular, homología y (pan)genómica microbiana

* tasas de evolución y selección de marcadores moleculares

Libros de referencia recomendados:

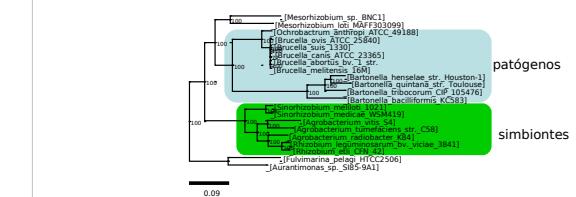
- Felsenstein, J. 2004. Inferring phylogenies. Sinauer Associates, INC., Sunderland, MA.
- Futuyma, D.J. 2017. Evolution, 4th Ed. Sinauer Associates, INC., Sunderland, MA.
- Graur, D., Li, W.H. 2000. Fundamentals of Molecular Evolution. Sinauer Associates, Inc., Sunderland.
- Lemey P., Salmin M., Vandamme A-M. 2010. The phylogenetic Handbook. Cambridge Univ. Press. UK
- Keith, Jonathan M. (Ed.) Bioinformatics: Volume I: Data, Sequence Analysis, and Evolution (Methods in Molecular Biology) <https://www.springer.com/gp/book/9781493966202>
- Nei, M., Kumar, S., 2000. Molecular Evolution and Phylogenetics. Oxford University Press, Inc., NY.
- Page, R.D.M., Holmes, E.C. 1998. Molecular Evolution - A Phylogenetic Approach. Blackwell Science Ltd, Oxford.
- Yang, Z., 2014. Molecular Evolution - A statistical approach. Oxford University Press, Oxford, UK

Introducción a la filoinformática – pan-genómica y filogenómica. TIB2019, Julio-Agosto 2019
<https://github.com/vinuesa/TIB-filoinfo>

La relación entre filogenética y evolución molecular:

- La filogenética tiene por objetivo trazar la relación ancestral descendiente de los organismos (árbol filogenético) a diferentes niveles taxonómicos, incluyendo el árbol universal, haciendo una reconstrucción de esta relación en base a diversos **caracteres homólogos**, tanto **morfológicos** como **moleculares**.

Las hipótesis filogenéticas resultantes son la base para hacer **predicciones** (inferencias) sobre propiedades biológicas de los grupos revelados por la filogenia mediante el mapeo de caracteres sobre la topología (hipótesis evolutiva). También proveen el contexto comparativo para poder inferir patrones de **evolución molecular**.



Evolución de la filogenética como disciplina científica



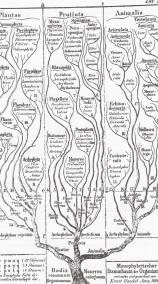
Los primeros intentos de reconstruir la historia filogenética estaban basados en pocos o ningún criterio objetivo.

Reflejaban las ideas o hipótesis plausibles generadas por expertos de grupos taxonómicos particulares.

La mayor parte de la 1a. mitad del SXIX los sistemáticos estaban más preocupados por el problema de definir a las especies biológicas, descubrir mecanismos de especiación y la variación geográfica de las especies, que en entender su filogenia.



No fue hasta los 40's y 50's que los esfuerzos de individuos como Walter Zimmermann y Willi Henning comenzaron a definir métodos objetivos para reconstruir filogenias en base a caracteres compartidos entre organismos fósiles y contemporáneos.



Filogenia y clasificación de la vida tal y como la propuso Ernst von Haeckel en 1866



Introducción a la Inferencia Filogenética – las especies y genomas microbianos

Introducción a la filoinformática – pan-genómica y filogenómica. TIB2019, Julio-Agosto 2019
<https://github.com/vinuesa/TIB-filoinfo>

¿Porqué estudiar filogenética y evolución molecular?

Corolario I:

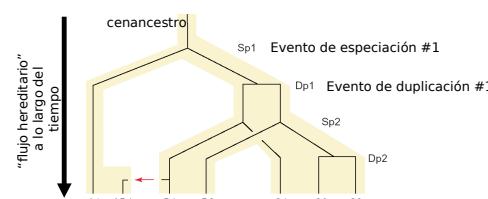
"Nothing in biology makes sense except in the light of evolution"
- Theodosius Dobzhanski, 1973
(The American Biology Teacher 35:125)

Corolario II:

"Nothing in evolutionary biology makes sense except in the light of a phylogeny"
- Jeff Palmer, Douglas Soltis, Mark Chase, 2004
(American J. Botany 91: 1437-1445)



El concepto de homología: definiciones básicas Subtipos de homología: ortología, paralogía y xenología



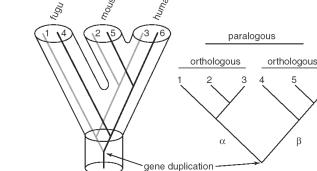
ortología: relación entre secuencias en la que la divergencia acontece tras un evento de especiación. El ancestro común es el cenácestro. La filogenia recuperada de estas secuencias refleja la filogenia de las especies.

paralogía: condición evolutiva en la que la divergencia observada acontece tras un evento de duplicación génica. La mezcla de ortólogos y parálogos en un mismo análisis filogenético recupera la filogenia correcta de los genes pero no necesariamente la de los organismos o taxa.

xenología: relación entre secuencias dada por un evento de transferencia horizontal entre linajes. Distorsiona fuertemente la filogenia de las especies.

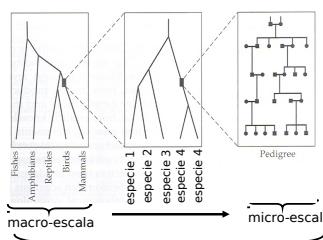
Arboles de genes vs. árboles de especies - el problema de la definición de relaciones de homología

- La **filogenia de especies** puede ser inferida erróneamente cuando se reconstruye en base a secuencias parálogas y no se muestran todas las copias (p. ej. si muestreamos sólo las copias 1, 3 y 5) eq ((fugu,human), mouse) !!!
- Más compleja aún es la estimación de la filogenia de especies si ha habido pérdida diferencial de parálogos en los díntos linajes a comparar
- Por tanto la inferencia de una filogenia de especies se realizará preferentemente usando genes de copia única, lo que hace más probable la condición de ortología



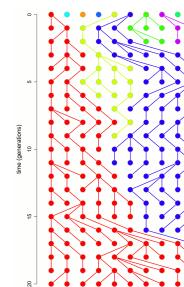
El concepto de filogenia y homología: definiciones básicas

"The stream of heredity makes phylogeny, in a sense, it is phylogeny. Complete genetic analysis would provide the most priceless data for the mapping of this stream".
G.G. Simpson (1945)

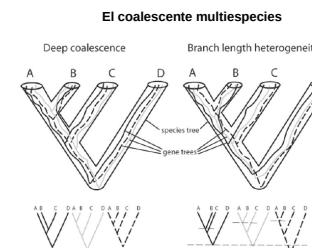


Filogenia: historia evolutiva del flujo hereditario a distintos niveles evolutivos/temporales, desde la genealogía de genes en poblaciones (micro-escala; dominio de la genética de poblaciones) hasta el árbol universal (macro-escala)

Filogenias de genes vs. filogenias de especies

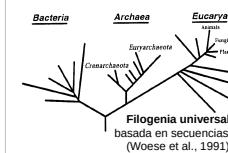


Modelo coalescente de genealogía de alelos en una población idealizada de Wright-Fisher evolucionando bajo deriva génica sin recombinación

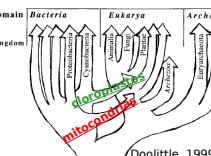


- Cada gen tiene su propia tasa de evolución
- Además pueden tener historias discordantes
- Por tanto, las filogenias de genes no reflejan necesariamente la filogenia de las especies

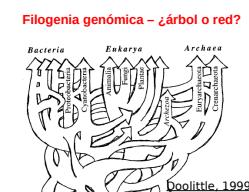
La TGH y la filogenia universal – problemas y limitaciones evidenciadas desde la perspectiva genómica



El origen evolutivo de los organelos: un caso clásico de TGH



Filogenia universal basada en secuencias rrs (Woese et al., 1991)

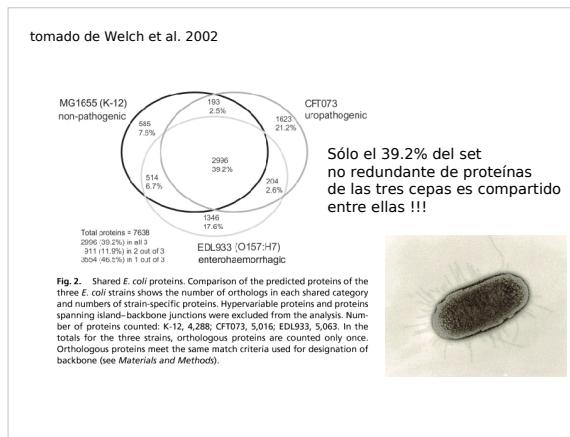
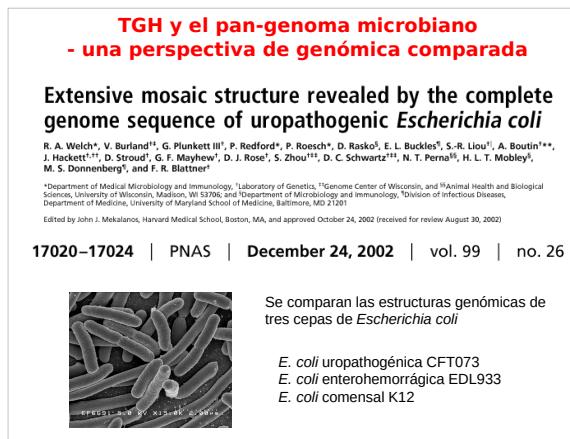


Problemas e incógnitas de las filogenias profundas:

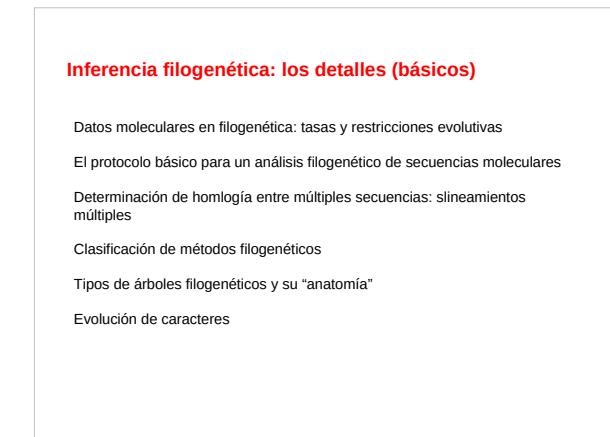
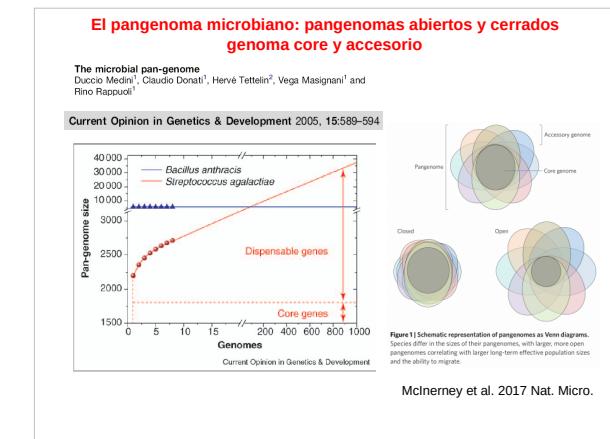
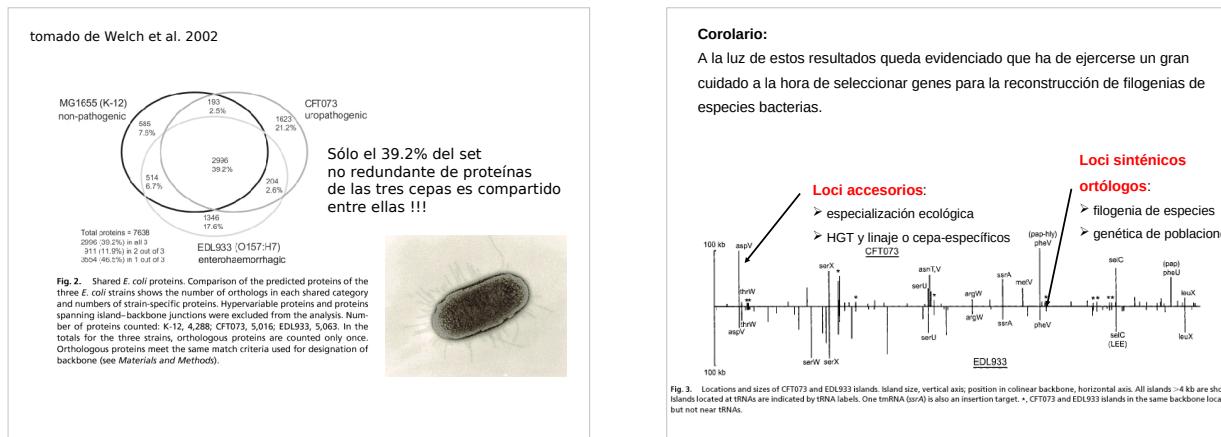
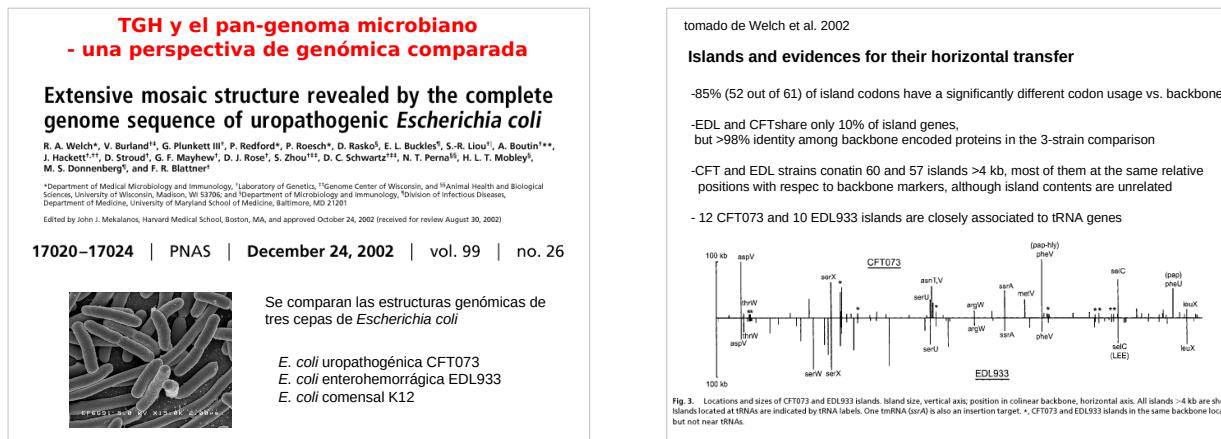
- ¿Cuántos ortólogos son realmente universales?
- Distinción entre ortólogos, parálogos y xenólogos
- Cuánta señal filogenética queda en las secuencias?
- ¿Alineamientos confiables?
- Métodos de reconstrucción y artefactos
- Congruencia de señales filogenéticas provenientes de distintos genes
- ¿Existió realmente un sólo ancestro?
- ... una larga lista

Introducción a la Inferencia Filogenética – las especies y genomas microbianos

Introducción a la filoinformática – pan-genómica y filogenómica. TIB2019, Julio-Agosto 2019
<https://github.com/vinuesa/TIB-filoinfo>



© Pablo Vinuesa 2019,
vinuesa{at}ccg{dot}unam{dot}mx
<http://www.ccg.unam.mx/~vinuesa/>



Introducción a la Inferencia Filogenética – las especies y genomas microbianos

Introducción a la filoinformática – pan-genómica y filogenómica. TIB2019, Julio-Agosto 2019
<https://github.com/vinuesa/TIB-filoinfo>

Aplicaciones y predicciones filogenéticas (II):
Evidencia molecular de transmisión de HIV-1 en un caso criminal usando genes de evol. rápida

Un gastroenterólogo fue acusado del intento de asesinato en 2º grado de su novia mediante inyección de sangre contaminada con HIV-1.

Este estudio representa el primer caso en el que reconstrucciones filogenéticas de secuencias (paciente P, víctima V y controles LA de portadores en la población) fueron admitidas en una corte criminal en EUA.

Las filogenias de RT y de env mostraron que las secuencias de la V compartían ancestría directa en forma de paralogía con las de una P del gastroenterólogo.

Análisis de posiciones de codones de la RT de la V revelaron genotipos consistentes con mutaciones que confieren AZTR, similares a las presentadas en la P.

El establecimiento a priori de la P y V como posible par de transmisión del HIV-1 representó una clara hipótesis para ser evaluada en marcos de estadística filogenética.

Ref: Metzker et al. 2002. PNAS 99:14292-142976

a 10% of bootstrap replicates place victim sequences within patient sequences

b Bootstrap replicates showing phylogenetic trees for patient (P), victim (V), and controls (LA).

Analisis de posiciones de codones de la RT de la V

a) Baylor College of Medicine, Houston, TX (BMC)

b) Dpt. Ecology and Evol. Biol., Univ. Michigan (MIC)

Homología entre secuencias de DNA: alineamientos múltiples

- A lo largo de la evolución las secuencias descendientes de otra ancestral van acumulando diversos tipos de mutaciones. Estas son **mutaciones puntuales** o **reorganizaciones genómicas**, que pueden involucrar **inserciones, deleciones, inversiones, translocaciones o duplicaciones**, mediados por distintos mecanismos de recombinación (homóloga e ilegítima)
- Cualquier análisis filogenético y/o evolutivo de secuencias moleculares requiere de **unalineamiento** para poder comparar sitios homólogos entre las secuencias a estudiar. Para ello se escriben las secuencias en filas una sobre la otra, de modo que los sitios homólogos quedan alineados por columnas. Cada sitio o columna del alineamiento corresponde a un carácter, y los nt o aa que ocupan dichas posiciones representan los distintosestados del carácter

Alineamiento múltiple de secuencias

Diagrama de alineamiento múltiple de secuencias de DNA. Se muestra un árbol filogenético y una matriz de alineamiento con colores que indican la presencia de bases específicas en cada posición.

Inferencia filogenética molecular - clasificación de métodos

Podemos clasificar a los métodos de reconstrucción filogenética en base al tipo de datos que emplean (**caracteres discretos vs. distancias**) y si usan un **método algorítmico** o un criterio de optimización para encontrar la topología

Tipo de datos

Método de reconstrucción	distancias	caracteres discretos
UPGMA	X	
Neighbour joining		X
Evolución mínima		X
Máxima parsimonia		X
Máxima verosimilitud		X

criterio de agrupamiento

criterio de optimización

Protocolo básico para un análisis filogenético de secuencias moleculares

colección de secuencias homólogas

- * BLAST y FASTA

alineamiento múltiple de secuencias

- Clustal, T-Coffee ...

análisis evolutivo del alineamiento y selección del modelo de sustitución más ajustado

- homogeneidad composicional, saturación
- selección de modelos, ...

estima filogenética

- NJ, ME, MP, ML, Bayes ...

pruebas de confiabilidad de la topología inferida

- proporciones de bootstrap probabilidad posterior ...

interpretación evolutiva y aplicación de las filogenias

Inferencia Filogenética - introducción

- La inferencia de relaciones filogenéticas a partir de secs. moleculares requiere de la selección de uno de los muchos métodos disponibles
- Con frecuencia la inferencia filogenética es considerada como una "caja negra" en la que "entran las secuencias y salen los árboles" (filogenias estimadas con MEGA)

Diagrama que muestra un fragmento de secuencia de DNA (GGCTTCAGTCATTCGG) y un alineamiento múltiple de secuencias. Una flecha apunta a un resultado de árbol filogenético:

```
sequences 1 2 3 4 5 6 7
Drosophila t t a t t a a
fugu a a t t t a a
mouse a a a a a a a
human a a a a a a a
```

distances

Drosophila	fugu	mouse	human
3	4	2	

parsimony

```
Drosophila 1 -> 2 -> 3 -> 4 -> 5 -> 6 -> mouse -> human
```

distance

```
Drosophila 2 -> mouse 1 -> human
```

• **Objetivos de esta taller son:**

- desarrollar un marco conceptual para entender los fundamentos teóricos (filosóficos) que distinguen a los distintos métodos de inferencia (clasificación de métodos)
- presentar el uso de **modelos y suposiciones** en filogenética
- manejo empírico de diversos paquetes de software para inferencia filogenética bajo diversos criterios

Métodos de reconstrucción filogenética – una clasificación

I.- **Tipos de datos: distancias vs. caracteres discretos**

- Los **métodos de distancia** primero convierten los alineamientos de secuencias en una matriz de distancias genéticas en base al modelo evolutivo seleccionado, la cual es usada por el método algorítmico de reconstrucción para calcular el árbol (UPGMA y NJ)
- Los **métodos discretos (MP, ML, Bayesianos)** consideran cada sitio del alineamiento (o una función probabilística para cada sitio) directamente

sequences 1 2 3 4 5 6 7

Drosophila	fugu	mouse	human
t t a t t a a	a a t t t a a	a a a a a a a	a a a a a a a

distances

Drosophila	fugu	mouse	human
3	4	2	

parsimony

```
Drosophila 1 -> 2 -> 3 -> 4 -> 5 -> 6 -> mouse -> human
```

distance

```
Drosophila 2 -> mouse 1 -> human
```

• Un set de 4 secs. y la matriz de distancias correspondiente

• Un árbol de parsimonia y uno de distancias para este set de datos produce topologías y longitudes de ramas idénticas

• La diferencia radica en que el árbol de parsimonia identifica qué sitio del alineamiento contribuye cada paso mutacional en la longitud de cada rama

Introducción a la Inferencia Filogenética – las especies y genomas microbianos

Introducción a la filoinformática – pan-genómica y filogenómica. TIB2019, Julio-Agosto 2019
<https://github.com/vinuesa/TIB-filoinfo>

Arboles filogenéticos: una introducción al bosque (I) terminología y conceptos básicos: anatomía de un árbol

- Definición: Un árbol filogenético es una estructura matemática usada para representar la historia evolutiva (relaciones de ancestro-descendiente) entre un grupo de secuencias o organismos. Dicho patrón de relaciones históricas es la estima hecha de la filogenia o árbol evolutivo.
- Anatomía básica de un árbol
 - A, B, C, D, E: nodos internos, vértice, grado 3
 - nodo terminal, hoja u OTU, grado 1
 - rama
 - split (bipartición) ($ABC|DE = \{*\} - \{*\}$)
 - nodo raíz, grado 2
- árboles no enraizados, sin direccionalidad
- árboles enraizados, con direccionalidad, que indica relaciones ancestro-descendiente (((humano, chimp), gorila), orang)
- reconstrucción de caracteres ancestrales
- longitud de ramas
- sopore o confianza en splits

Arboles filogenéticos: una introducción al bosque (III) terminología y conceptos básicos

- Los árboles son como móviles: las ramas pueden rotarse sobre sí mismas sin afectar a las relaciones entre los OTUs; (((A,B),C),D),E) se puede representar como:
- Los árboles presentan distintos grados de resolución
 - topología estrella
 - topología parcialmente resuelta
 - topología totalmente resuelta
 - politomías

Arboles filogenéticos: una introducción al bosque (V) terminología y conceptos básicos

- Terminología relacionada con la reconstrucción de la historia de cambios en estados de carácter
 - apomorfia: carácter derivado; estado apomórfico
 - plesiomorfia: caract. ancestral; estado plesiomórfico o ancestral
- autapomorfía
- sinapomorfía
- homoplasia
- carácter derivado único (aut)
- carácter derivado compartido (syn)
- carácter compartido no homólogo, es decir, no heredado directamente del ancestro

Arboles filogenéticos: una introducción al bosque (II) enraizamiento de árboles

- Tres métodos usados para el enraizado de árboles:
 - grupo externo - (invertebrado) a grupo interno (vertebrados)
 - punto medio – se pone la raíz en el punto intermedio del camino más largo del árbol
 - duplicación génica – enraizamos en el nodo que separa a las copias parálogas
- La mayoría de los métodos de reconstrucción estiman árboles no enraizados, por lo que no discernen entre las 5 posibles topologías enraizadas generables a partir de 4 OTUs.
- Para enraizar un árbol (decidir cuál topología es la que refleja el proceso evolutivo), necesitamos información biológica adicional

Arboles filogenéticos: una introducción al bosque (IV) terminología y conceptos básicos: tipos de árboles

- Un cladograma: sólo indica las relaciones de ancestría entre OTUs
- Una topología aditiva contiene la información sobre longitudes de ramas, que refleja la distancia genética entre OTUs. Así entre R. galegae y R. huautlense la distancia estimada es de: $0.05 + 0.06 = 0.11$
- Una topología ultramétrica, dendrograma o árbol linealizado, representa un tipo especial de árbol aditivo en el que los nodos terminales son todas equidistantes de la raíz. Este tipo de árbol se emplea para representar el tiempo evolutivo, expresado bien como años o cantidad de divergencia medida por un reloj molecular

Arboles filogenéticos: una introducción al bosque (VI) terminología y conceptos básicos

- Tipos de homoplasia
 - evolución paralela
 - evolución convergente
 - pérdida secundaria
- Reversión a la condición ancestral
- Evolución independiente del mismo estado de carácter a partir de la misma condición ancestral
- Evolución independiente del mismo estado de carácter a partir de una condición ancestral diferente

Introducción a la Inferencia Filogenética – las especies y genomas microbianos

Introducción a la filoinformática – pan-genómica y filogenómica. TIB2019, Julio-Agosto 2019
<https://github.com/vinuesa/TIB-filoinfo>

